Temporal Decomposition と STRAIGHT を用いた 低ビットレート音声符号化に関する研究

越智 崇夫 (010029)

北陸先端科学技術大学院大学 情報科学研究科

2002 年 2 月 15 日

キーワード: Temporal Decomposition, STRAIGHT, ベクトル量子化, ビットレート.

1 はじめに

近年の携帯電話やマルチメディア通信の発達と普及率の増加に伴い、能率的な伝送または記録を行うことができる音声符号化の需要が高まっている。

より低ビットレートの音声符号化システムを構築するためには、音声学的情報を上手く捉えて符号化する必要があり、何が音声情報をよく特徴づけているかということが問題となる。現在、様々な手法を用いた低ビットレートの音声符号化の研究がが行われているが、2 kbit/s 以下のビットレートでは十分な品質の符号化システムは実現されていない。

本研究では、合成音の品質を向上させるために、音声分析・変換・合成方式として高音質な合成音を作成することができる Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram (STRAIGHT)[1] を用いる。しかし、符号化システムとしては、かなり多くの情報を伝送することになり、情報圧縮という点では不利である。そこで、STRAIGHT により音声データからスペクトル情報と基本周波数情報を抽出した後に、Temporal Decomposition（TD）を用いて音声信号の時間的な変動に極在して現れる音声学的情報を分解する。分解することによって、より低ビットな特徴づけを目指す。さらに、それを基にした低ビットレート音声符号化システムを構築する。

2 符号化システム

図 1 にシステムの概要を示す。STRAIGHT において、合成側に送られる情報は、基本周波数情報（F_0）と平滑化されたスペクトル情報である。スペクトル情報を LSF に変換し、Modified Restricted Temporal Decomposition (MRTD)[3] を用いてスペクトルパラメータの時間変化パターン（イベント関数）とスペクトルの安定する位置におけるスペクトル情報（イベントターゲット）に分解する。分解したパラメータをベクトル量子化することに

より、スペクトル情報を圧縮する。その他のパラメータに対しては、スカラー量子化を適用する。

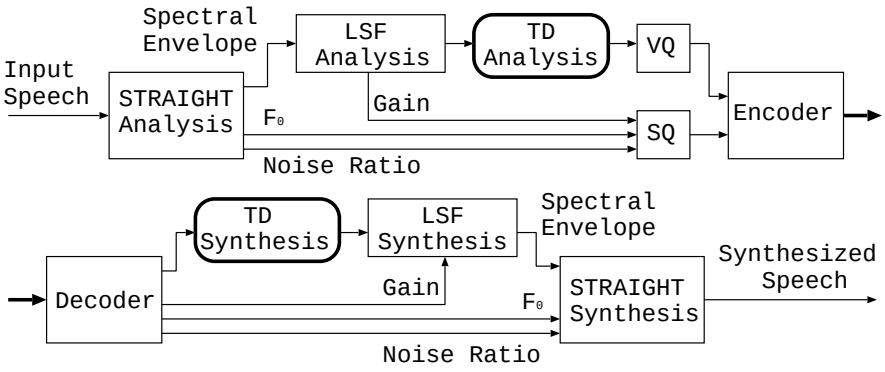

図 1: TD と STRAIGHT を用いた低ビットレート音声符号化システム

3 STRAIGHT

音声符号化における分析合成符号化方式は、音声生成モデルに基づいて符号化を行うことにより高い情報圧縮率を実現しているが、高品質な合成音を得ることができないという欠点がある。しかし、河原らによって提案された音声分析・変換・合成方式 STRAIGHT は、分析合成方式ながら高品質な合成音を得られる方法として注目を浴びている。。

そこで、本研究では STRAIGHT を用いることによって、合成音の品質改善を図る。

4 スペクトル情報の符号化

4.1 LSF

STRAIGHT で得られる振幅スペクトル $X[k]$、$0 \leq k \leq N/2$ を用いてパワースペクトル $S[k]$ を計算する。

$$S[k] = |X[k]|^2, \quad 0 \leq k \leq N/2$$

パワースペクトルからフーリエ逆変換することによって相関関数を求めると次のようになる。

$$R[n] = \frac{1}{N} \sum_{k=0}^{N-1} S[k] \exp\left\{ j \frac{2\pi kn}{N} \right\}$$

ここで、$S[k] = S[N-k]$。この相関関数を有する過程 $x(n)$ が全極フィルタ（次数 L）からの出力と仮定すれば、フィルタの係数を a_l^L、$l = 1, 2, ..., L$、$0 < L < N/2$ として、

$$P_L = R[0] - \sum_{l=1}^{L} a_l^L R[l]$$

と書ける。ここで、P_L は誤差（ゲイン）である。P_L が最小となるようにフィルタの係数 a_l^L を決定する。このときのフィルタの係数は、LPC の予測係数と一致する。予測係数 a_l^L を用いて LSF を計算する。

4.2 制限と修正を加えた時間分解法 [3]

LSF に変換されたスペクトル情報は、さらに MRTD[3] を用いてイベント関数とイベントターゲットに分解される。MRTD は、より低ビットレート音声符号化に適応するため、TD[2] に制限と修正を加えた手法である。TD[2] は以下のように、イベントベクトルの線形結合によってスペクトルパラメータの時間変化を近似する。

$$\hat{\boldsymbol{y}}(n) = \sum_{k=1}^{K} \boldsymbol{a}_k \phi_k(n), \quad 1 \le n \le N \tag{1}$$

ここで、$\boldsymbol{a}_k$、$\phi_k(n)$ は、それぞれ k 番目イベントターゲット、イベント関数である。$\hat{\boldsymbol{y}}(n)$ は、n 番目スペクトルパラメータ $\boldsymbol{y}(n)$ の近似値である。

MRTD では、イベント関数に 2 つの制約が加えられる。1）時間のどの瞬間においても、隣接する 2 つのイベント関数だけで記述する。2）どの時刻においても隣接するイベント関数の合計は 1 である。この制約を用いれば式（1）は次のようになる。$C(k) \le n \le C(k+1)$ に対して

$$\begin{aligned} \hat{\boldsymbol{y}}(n) &= \boldsymbol{a}_k \phi_k(n) + \boldsymbol{a}_{k+1} \phi_{k+1}(n) \\ &= \boldsymbol{a}_k \phi_k(n) + \boldsymbol{a}_{k+1}(1 - \phi_k(n)) \end{aligned}$$

ここで、$C(k)$、$C(k+1)$ は、それぞれイベント k、$k+1$ の中心位置である。ただし、

$$\begin{gathered} \phi_k(C(k)) = 1, \ \phi_k(C(k+1)) = 0 \\ 0 \le \phi_k(n) \le 1 \ \text{ for } \ C(k) \le n \le C(k+1) \end{gathered}$$

最終的に $\phi_k(n)$ は、次のように決定される。

$$\phi_k(n) = \begin{cases} 1 - \phi_{k-1}(n), & \text{if } C(k-1) < n < C(k) \\ 1, & \text{if } n = C(k) \\ \min(\phi_k(n-1), & \\ \max(0, \hat{\phi}_k(n))), & \text{if } C(k) < n < C(k+1) \\ 0, & \text{その他} \end{cases}$$

ここで

$$\hat{\phi}_k = \frac{\langle (\boldsymbol{y}(n) - \boldsymbol{a}_{k+1}), (\boldsymbol{a}_k - \boldsymbol{a}_{k+1}) \rangle}{\| \boldsymbol{a}_k - \boldsymbol{a}_{k+1}) \|^2}$$

4.3 LSF の次数決定

4.3.1 次数に対するスペクトル歪みの変動

LSF の次数を決定するために、符号化システムにおける対数スペクトル歪みを調べた。テストデータとして、ATR 日本語音声データベースの話者 MMY による音韻バランス 503 文章中の 112 文を 8 kHz にダウンサンプリングしたものを用いた。スペクトル情報の補間方法に LSF のみを適用した場合、LSF および MRTD を適用した場合の結果を図 2 に示す。ただし、量子化は行っていない。横軸は LSF の次数を表し、縦軸は対数スペクトル歪みを表す。図 2 より、次数を 22 以上にしても LSF および MRTD 後のスペクトル歪みの著しい改善は期待できないことがわかる。

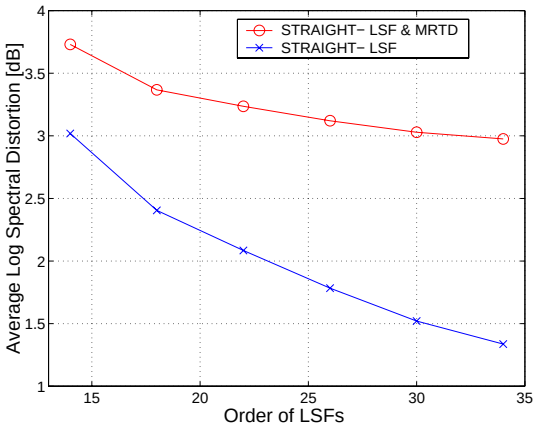

図 2: LSF の次数に対するスペクトル歪みの変動

4.3.2 次数に対する音声品質の変動

LSF の次数を変化させた場合における合成音の品質を、聴取実験 (シェッフェの一対比較法) により主観的に評価した。約 2 秒間ごとに異なる刺激音を一対として呈示し、どちらの音（前者・後者）の歪みが小さいかを 5 段階で判断させた。被験者は正常聴力を有すると認められる大学院生 6 名とした。聴取実験には、ATR 音声データベースの話者 MMY による音韻バランス 503 文章中の 2 文章を用いた。データは、8 kHz にダウンサンプリングしたものを用いた。この 2 文章に対して、LSF の次数を 10、14、18、22、26、30 と変化させたものに MRTD を適用して分析合成を行った。ただし、分析合成を行う際に量子化は行っていない。実験結果を図 3 に示す。横軸は母数を表し、その位置は呈示した刺激音の相対的な距離を表す。プラス側（右側）にいくほど歪みが小さく、マイナス側（左側）にいくほど歪みが大きいと判断される。矢印の上の数字は、LSF の次数を表す。実験より、LSF 次数を 22 以上にしても聴覚的に歪みの改善は感じられないことが示された。よって、LSF の次数を 22 次に決定した。

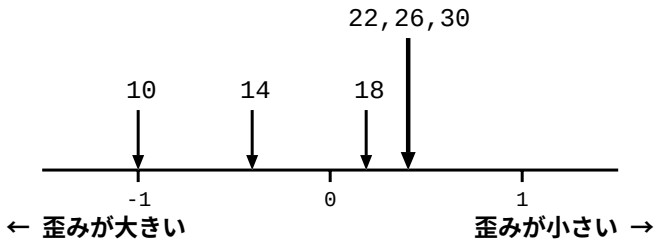

図 3: LSF の次数に対する音声品質の変動

4.4 ベクトル量子化

MRTD によって得られたイベント関数とイベントターゲットの量子化方法には、ベクトル量子化を用いる。

イベント関数の量子化
イベント関数の時間長は、各イベントごとに異なるため、その時間長を 16 次のベクトルに正規化して 7 bit でベクトル量子化を行った。

イベントターゲットの量子化
イベントターゲットは、分割ベクトル量子化を行う。低次の LSF 値の分散が大きいことから 6、7、9 次に 3 分割して、各ベクトルごとに 8 bit または 9 bit 割り当てた。それぞれ 1 イベントに割り当てるビット数は合計で、24、27 bit となる。

5 基本周波数の符号化

STRAIGHT により求めた基本周波数から、無声区間と有声区間の時間長を量子化する。有声部分に対しては、データを 28 ms ごとに取り、5 bit で対数スカラー量子化を行った。合成側において、これらの情報をもとに無声区間では 0 として、有声区間では線形補間を用いて元の長さに復元し、基本周波数を再構成した。RMS 誤差は、約 3.7 Hz であった。

6 ゲインの符号化

ゲインは、20 ms ごとにサンプリングし、6 bit で対数スカラー量子化を行う。量子化されたパラメータは、合成側でスプライン補間を用いて元のサイズに復元される。前節と同様のデータを用いた場合に、RMS 誤差は約 3.5 dB であった。

7　雑音比の符号化

雑音比パラメータは、雑音比ターゲットとスペクトル用のイベント関数を用いて次のように再構成できる。

$$\hat{i}(n) = \sum_{k=1}^{K} i_k \phi_k(n), \quad 1 \leq n \leq N$$

ここで、$\hat{i}(n)$ と i_k は、それぞれ n 番目のフレームに対して再構成した雑音比パラメータと雑音比ターゲットである。雑音比ターゲットは次のように、元の雑音比パラメータと再構成した雑音比パラメータの二乗誤差を最小にするように決定される。雑音比ターゲットは、5 bit でスカラー量子化した。RMS 誤差は 0.1 であった。

$$E_i = \sum_{n=1}^{N} \Big(i(n) - \sum_{k=1}^{K} i_k \phi_k(n) \Big)^2$$

8　提案法のビット割り当て

各パラメータのビット割り当てを表 1 に示す。ただし、イベント数は約 15 events/s になるように設定した。イベントターゲットに対するビット割り当ての括弧内の値は、分割したイベントターゲットにそれぞれ割り当てたビット数である。

表 1: ビット割り当て

パラメータ	提案法 1	提案法 2
イベントターゲット	24(8+8+8)	27(9+9+9)
イベント関数	7	7
イベント間の距離	8	8
雑音比	5	5
小計 A (合計×イベント数)	660 [bit/s]	705 [bit/s]
基本周波数	215	215
ゲイン	300	300
入力音声の最大値	9	9
小計 B	524 [bit/s]	524 [bit/s]
総計 (A+B)	1194 [bit/s]	1229 [bit/s]

9　品質評価実験

提案法の品質を評価するために、提案法による合成音と、他の低ビットレート音声符号化方式による合成音との品質比較実験を行った。

実験はシェッフェの一対比較法により行った。約 2 秒間ごとに異なる刺激音を一対として呈示し、どちらの音（前者・後者）の歪みが小さいかを 5 段階で判断させた。被験者は正常聴力を有すると認められる大学院生 8 名とした。符合帳の学習データは、ATR 日本語

音声データベースにおける音韻バランス 503 文章中の約 108 文章を用いた。ただし、データは 8kHz にダウンサンプリングしたものを用いている。話者は男女各 3 名である。音声データは、学習外男女各 1 名の発話音声 2 文章（学習外）を用いた。この各データに対して、STRAIGHT に LSF を適用したもの（量子化なし）、STRAIGHT に LSF と MRTD を適用したもの（量子化なし）、4.8 kbit/s CELP、2.4 kbit/s LPC10、1.19 kbit/s に設定した提案法、1.23 kbit/s に設定した提案法の 6 つの方法によって刺激音を作成した。

品質評価実験を行った結果、4.8 kbit/s CELP の品質には及ばないものの、2.4 kbit/s LPC10 よりも明らかに良い品質を持っていることがわかった。

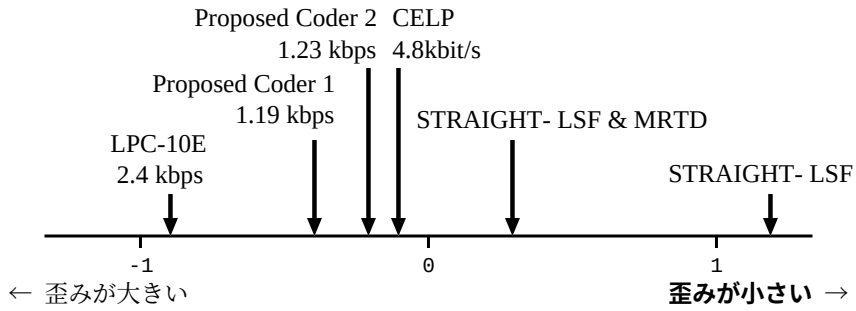

図 4: 品質評価実験の結果

10 まとめ

本研究では、TD およびベクトル量子化を用いることで、STRAIGHT で求めたスペクトル情報を圧縮した。その他のパラメターはスカラー量子化を行った。最終的にそれらを基にした約 1.2 kbit/s の低ビットレート音声符号化システムを構築し、聴取実験により品質評価を行った。

品質評価実験を行った結果、4.8 kbit/s CELP の品質には及ばないものの、2.4 kbit/s LPC10 よりも明らかに良い品質を持っていることがわかった。よって、提案法は、低ビットレート音声符号化において高品質ではないが、2 kbit/s 以下のビットレートでも十分な品質の合成音を作成できる可能性があると言える。

参考文献

[1] 河原英紀, “聴覚の情景分析と高品質音声分析変換合成法 STRAIGHT,” 音響学会講演論文集, 1-2-1, pp.186-192, 1997-9.

[2] B.S.Atal, “Efficient coding of LPC parameters by temporal decomposition,” Proc. ICASSP ’83, pp.81-84, 1983.

[3] P.C.Nguyen and M.Akagi, "Improvement of the restricted temporal decomposition method for LSF parameters," Proc. 2001 Autumn Meeting of ASJ, pp.267-268, 2001.